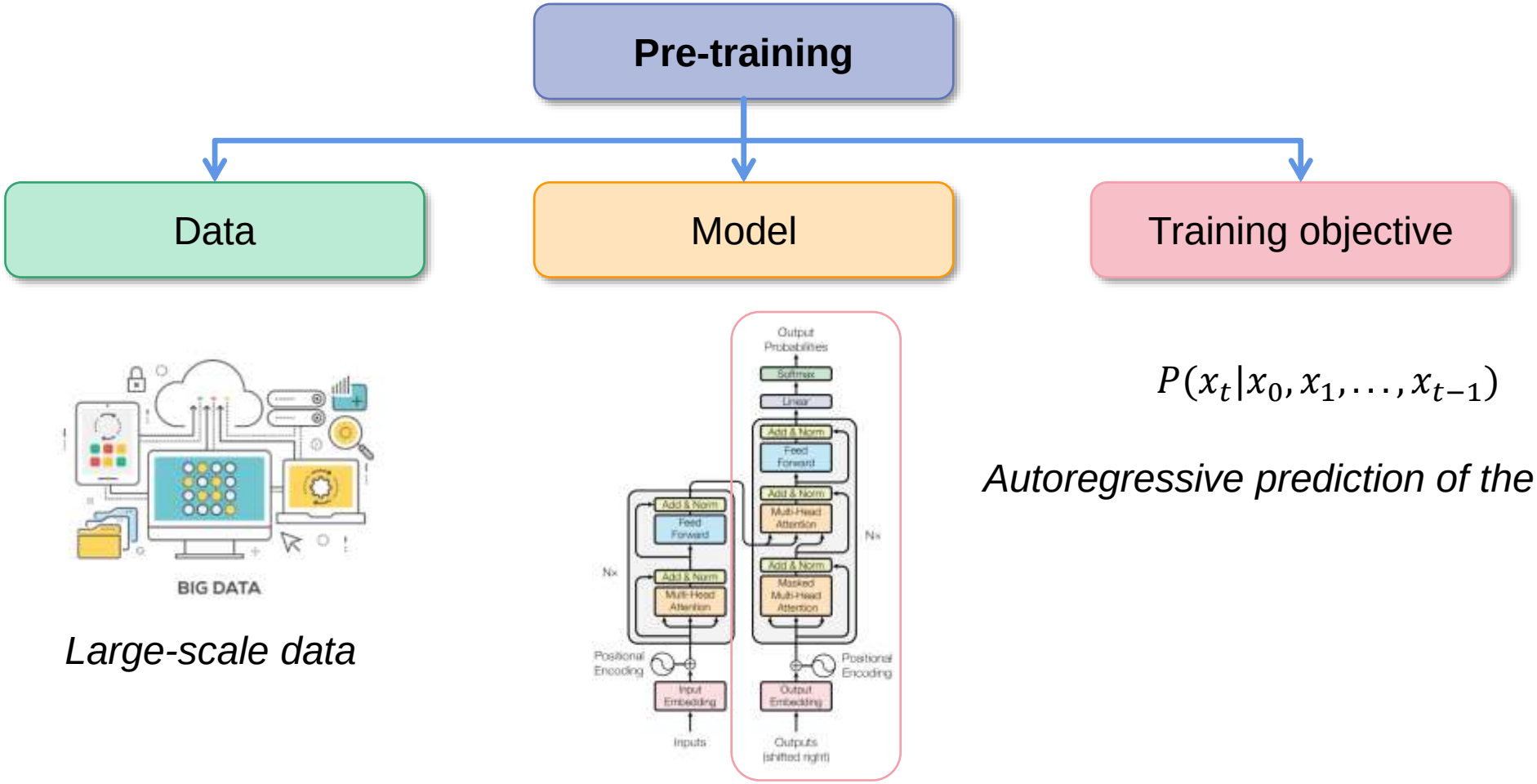
RL in LLMs

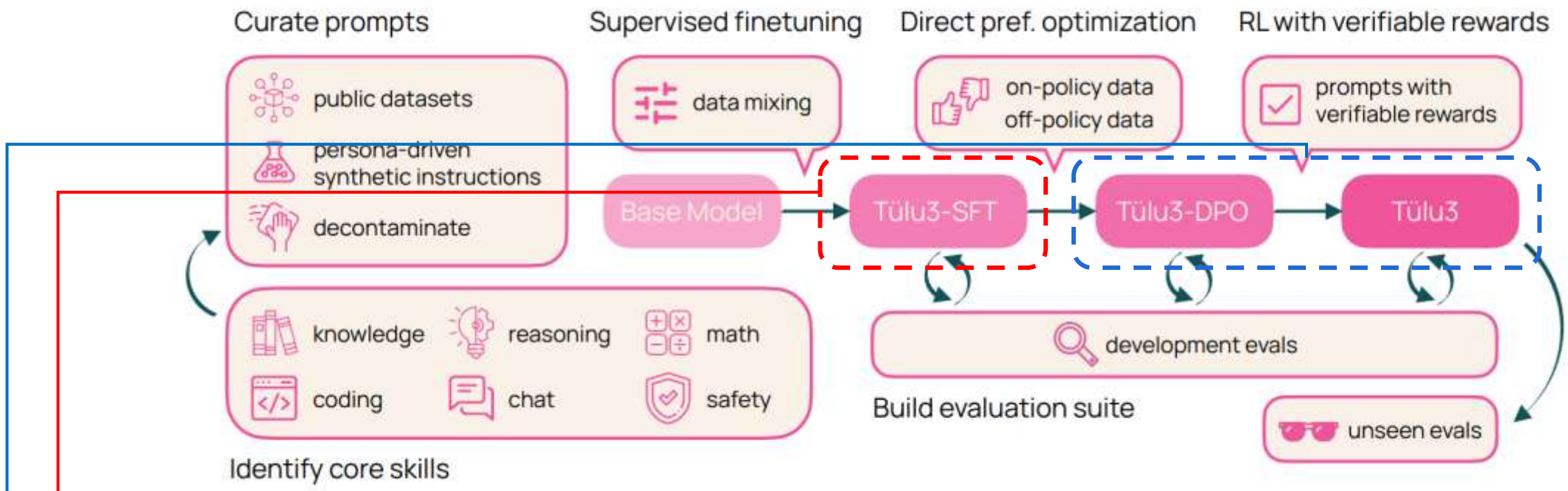
Pre-training



✓ Build a **base model** with general language understanding and generation capabilities.

Post-training

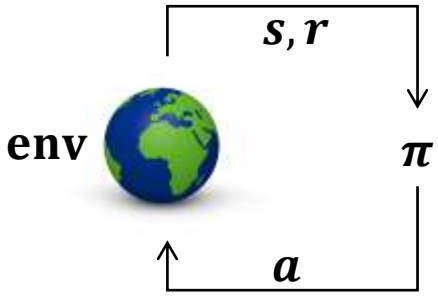
Tülu3



1. **Supervised Fine-Tuing (SFT)**: use data with annotations to build the ability of *instruction following*
2. **Reinforcement Learning (RL)**:
   - optimize the model behavior to *align with human preferences*
   - promote the evolution of model *autonomous reasoning ability*

Background - RL

Concept



- env → s_0
- $\pi \rightarrow a_0 \sim \pi(\cdot|s_0)$
- env → s_1, r_0
- ...

MDP(Markov Decision Process)

Definition: (S, A, R, P, γ)

$s \in S \quad a \in A \quad r_t = R(s_t, a_t)$

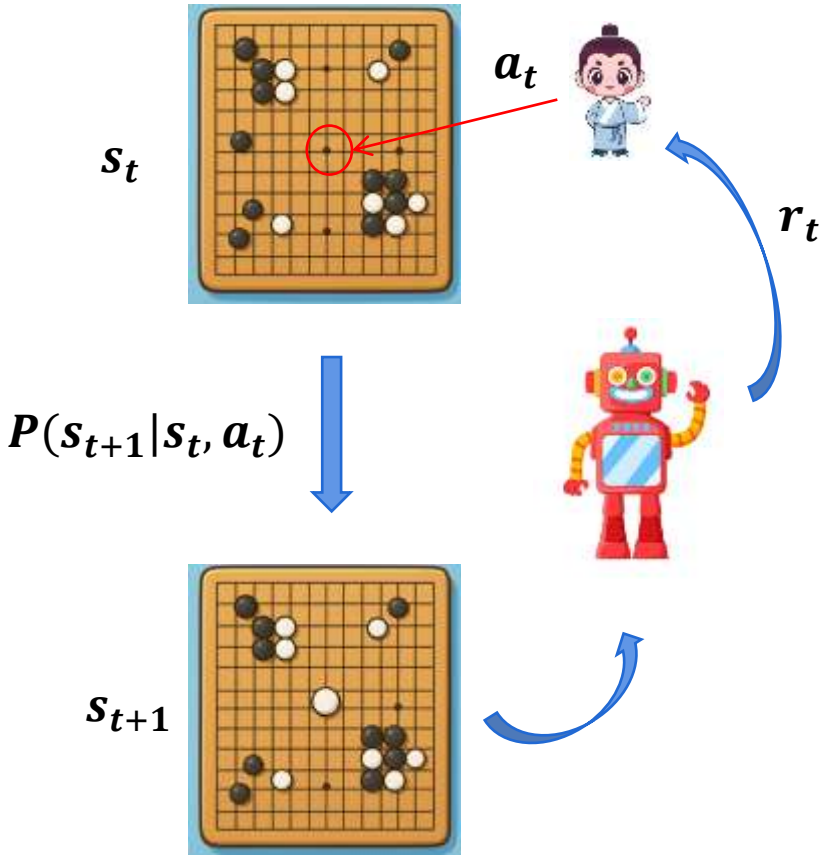
Markov property: $P(s_{t+1}|s_{<t}, a_{<t}, s_t, a_t)$

$P(s_{t+1}|s_t, a_t)$

Policy objective:

$$J(\pi) = \mathbb{E}_{a_0 \sim \pi(\cdot|s_0)}\left[r(s_0, a_0) + \gamma \mathbb{E}_{s_1 \sim P(\cdot|s_0, a_0)}\left[\mathbb{E}_{a_1 \sim \pi(\cdot|s_1)}[r(s_1, a_1) + \cdots]\right]\right]$$

Example



RL for LLMs

MDP: (S, A, R, P, γ) ⟹ Language-augmented MDP: (V, S, A, R, P, γ)

V: vocabulary A specific token: $w \in V$

S: $S \subset V^M$ $s = (w_1, w_2, w_3, \ldots, w_M)$ s_0: prompt

A: $S \subset V^N$ $a = (w_1, w_2, \ldots, w_N)$

R: $r = R(s, a)$

P: $S \times V \to S$ $s_{i+1} = (s_i, w_i) = (s_0, w_{1:i+1})$ auto-regressive paradigm

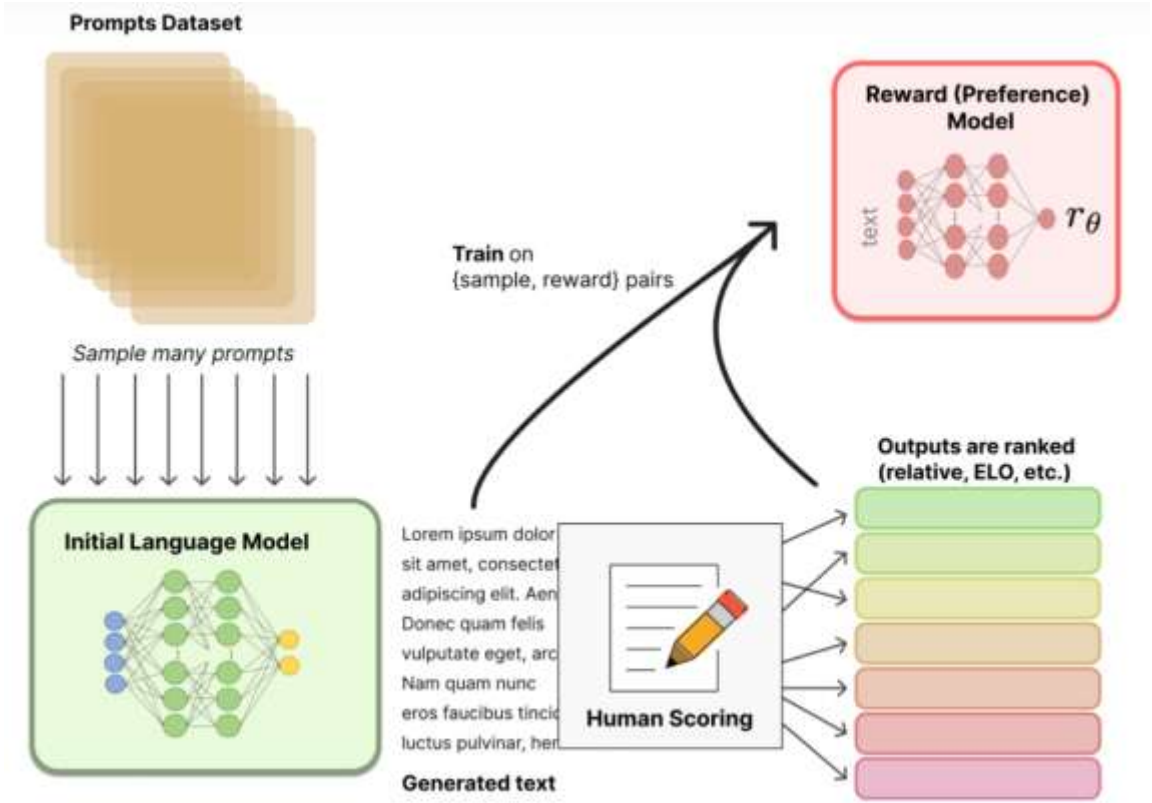
Token-level policy: $\pi(w_i | s_0, w_{1:i-1})$

Sentence-level policy: $\pi(a|s_0) = \prod_{i=1}^{N} \pi(w_i | s_0, w_{1:i-1})$

Policy objective:

$$J(\pi) = \mathbb{E}_{s_0 \sim D}\left[\mathbb{E}_{a \sim \pi(\cdot|s_0)}[r(s_0, a)]\right]$$

Why use RL?



Prompt: s
Candidate answers: a_1, a_2, a_3, a_4
Quality rank: $a_1 > a_2 > a_3 > a_4$

How can we optmize the model?

Supervised learning: $\mathbf{max}\ \mathbf{log}\ \boldsymbol{p}(\boldsymbol{a_1}|\boldsymbol{s})$

⇩

Data waste & poor generalization

A solution:

for each answers: $\mathbf{max}\ \ \boldsymbol{R_i} \cdot \mathbf{log}\ \boldsymbol{p}(\boldsymbol{a_i}|\boldsymbol{s})$

gradient: $\boldsymbol{R_i} \cdot \boldsymbol{\nabla_\theta} \mathbf{log}\ \boldsymbol{p_\theta}(\boldsymbol{a_i}|\boldsymbol{s})$

↓

✓ **The policy gradient of RL**

Policy Gradient

The goal of policy in RL: $J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$

The policy gradient: $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^{T} \nabla_\theta \log\pi_\theta(a_t|s_t) R_t]$, where $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$

Single Step $\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s_0\sim D, a\sim\pi_\theta(\cdot|s_0)}[\boldsymbol{r(s_0,a)\nabla_\theta \log \pi_\theta(a|s_0)}]$ $\longleftrightarrow$ $\boldsymbol{R_i \cdot \nabla_\theta \log p_\theta(a_i|s)}$

or $\boldsymbol{A(s_0,a)} = r(s_0,a) - b$

The goal of RL in LLMs is to fine-tune the pretrained model to maximize the probability of high-quality responses

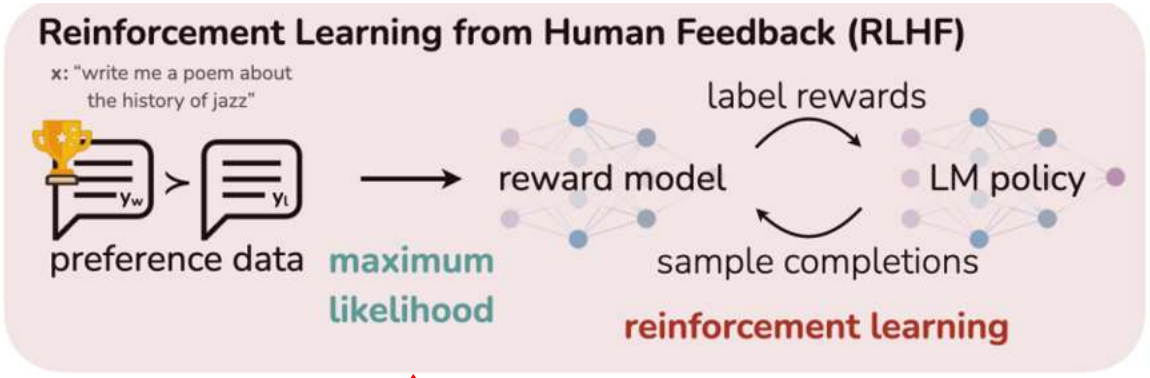
$$J(\pi_\theta) = \mathbb{E}_{s_0\sim D}\left[\mathbb{E}_{a\sim\pi_\theta(\cdot|s_0)}[r(s_0,a)] - \boldsymbol{\beta D_{KL}(\pi_\theta, \pi_{ref})}\right]$$

Sentence-level $R(s_0,a) = r(s_0,a) - \beta \log \frac{\pi_\theta(a|s_0)}{\pi_{ref}(a|s_0)}$

token-level

$$R(s_0,w_i) = \begin{cases} -\beta \sum_{i=1}^{T} \log \frac{\pi_\theta(w_i|s_0)}{\pi_{ref}(w_i|s_0)} & T < N \\ r(s_0,a) - \sum_{i=1}^{N} \log \frac{\pi_\theta(w_i|s_0)}{\pi_{ref}(w_i|s_0)} & T = N \end{cases}$$

The Evolution of RL in LLMs



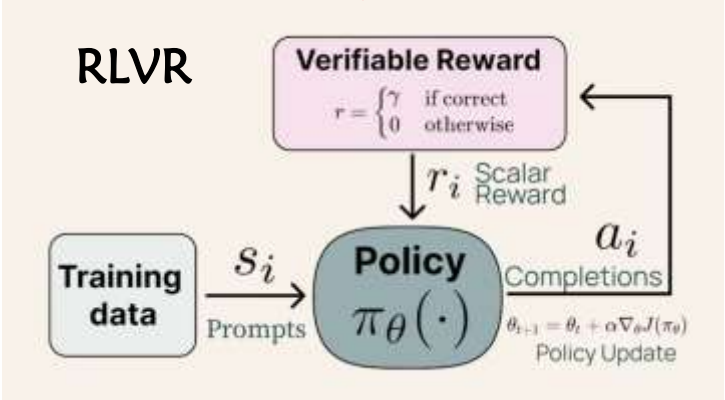
Optimize model to **align human preference** based on the answers with **manually annotated preferences**

?

GPT-4/Deepseek-R1
2024

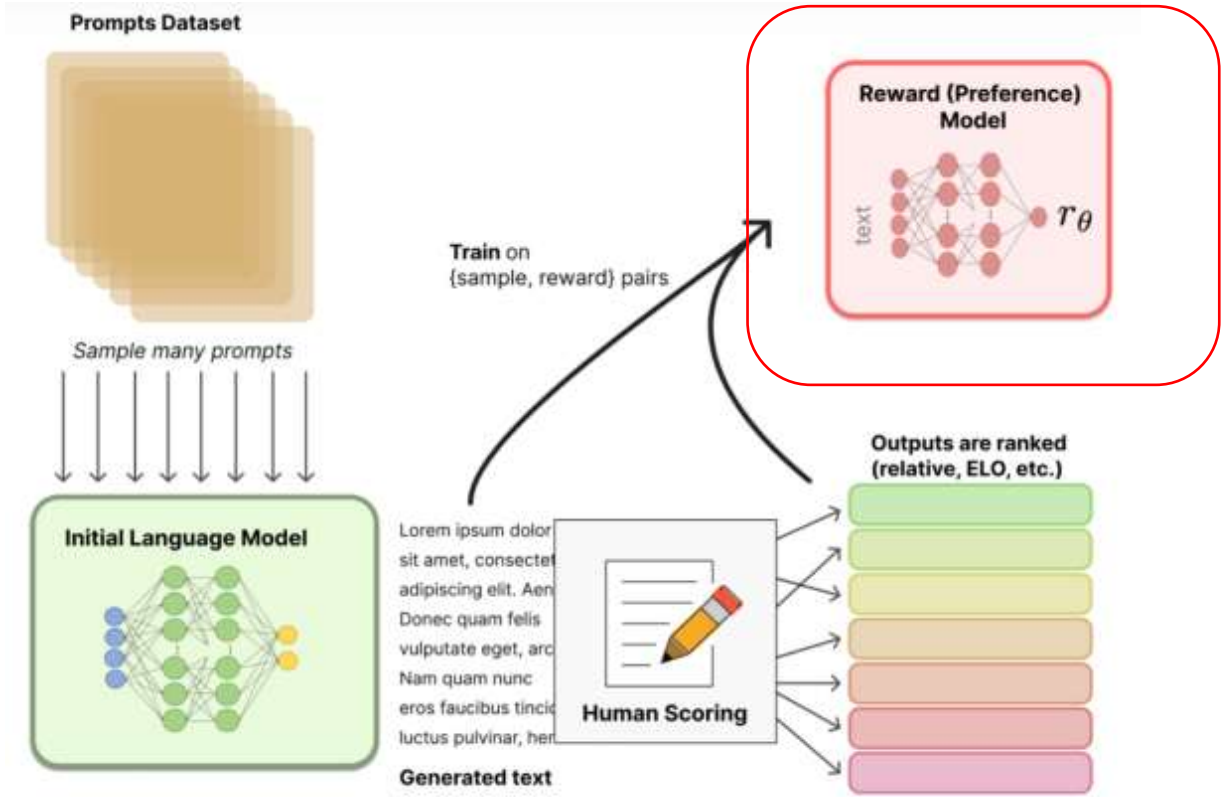
InstructGPT/ChatGPT
2022

Timeline



Automatically **verify the answers** to obtain rewards

The First Stage - RLHF



Preference-based reward:

$$\hat{P}(a_c > a_r|s) = \frac{exp[\hat{r}(s,a_c)]}{\exp[\hat{r}(s,a_c)] + exp[\hat{r}(s,a_r)]}$$ Bradley-Terry (BT) model

CE loss

$$L(\hat{r}) = -\mathbb{E}_{(s,a_c,a_r)\sim D}[\log \hat{P}(a_c > a_r|s)]$$

$$= -\mathbb{E}_{(s,a_c,a_r)\sim D}[\log(\sigma(\hat{r}(s,a_c) - \hat{r}(s,a_r)))]$$

Algorithm - PPO

$$\max_{\pi_\theta} \mathbb{E}_s \, \mathbb{E}_{a\sim\pi_{old}(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)} A^{\pi_{old}}(s,a) \right]$$

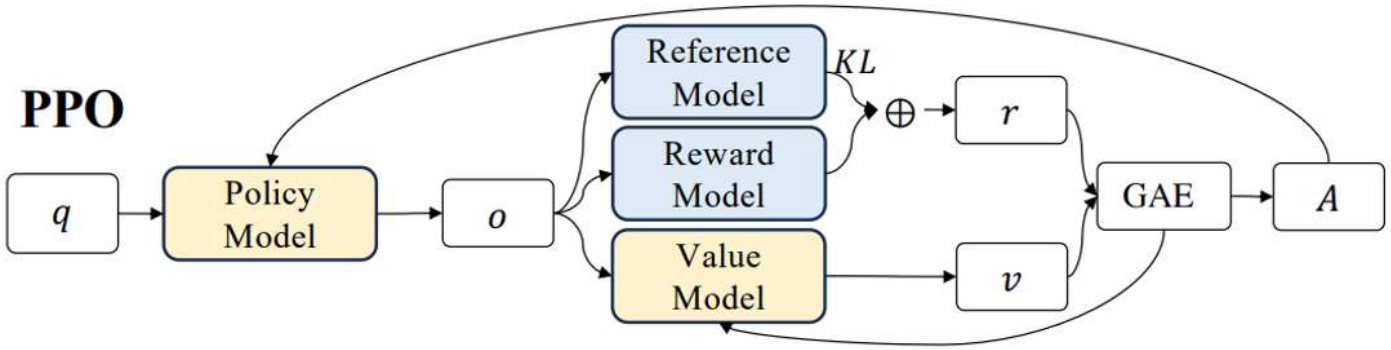
importance sampling

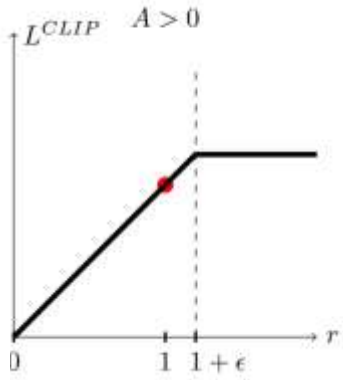
estimated based on a value model

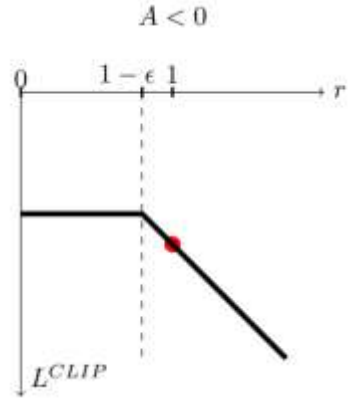
$$\text{s.t.} \quad \mathbb{E}_s[D_{KL}(\pi_{old}(a|s), \pi_\theta(\cdot|s))] < \delta$$

$$\operatorname{argmax}_\theta \mathbb{E}_s \, \mathbb{E}_{a\sim\pi_{old}(\cdot|s)} \left[\min \left(\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)} A^{\pi_{old}}(s,a), \text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{old}(a|s)}, 1-\epsilon, 1+\epsilon \right) A^{\pi_{old}}(s,a) \right) \right]$$

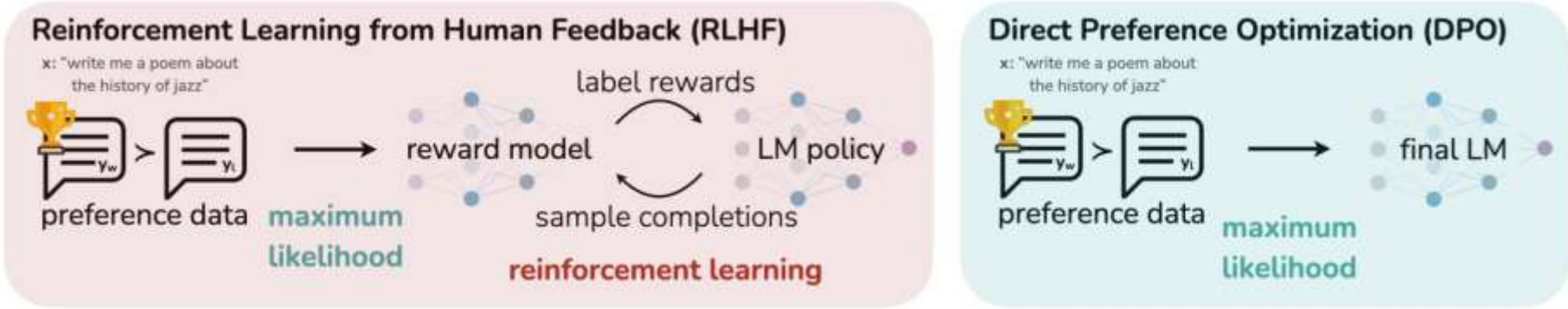
✓ **restricts how much the model policy can change at each step**







Algorithm - DPO



- rewrite the objective

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi}\left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi(y|x)\,||\,\pi_{\mathrm{ref}}(y|x)\right]$$

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right] \quad Z(x) = \sum_{y}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

- explicit reward

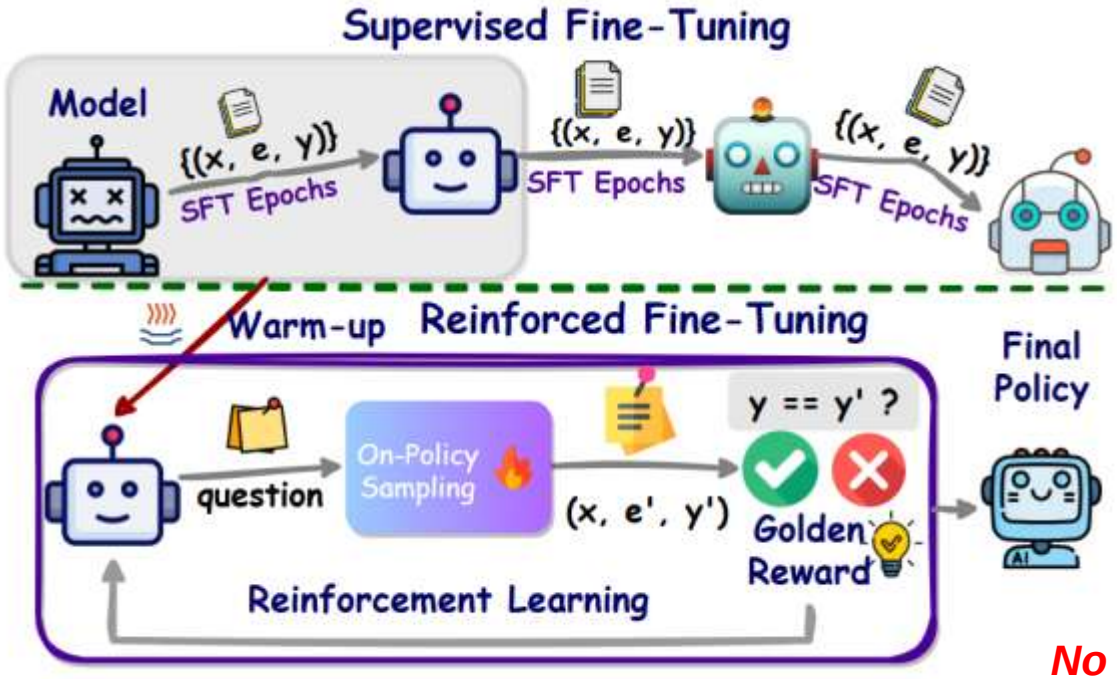
$$\pi^*(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right) \rightarrow r^*(x,y) = \beta\log\frac{\pi^*(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta\log Z(x)$$

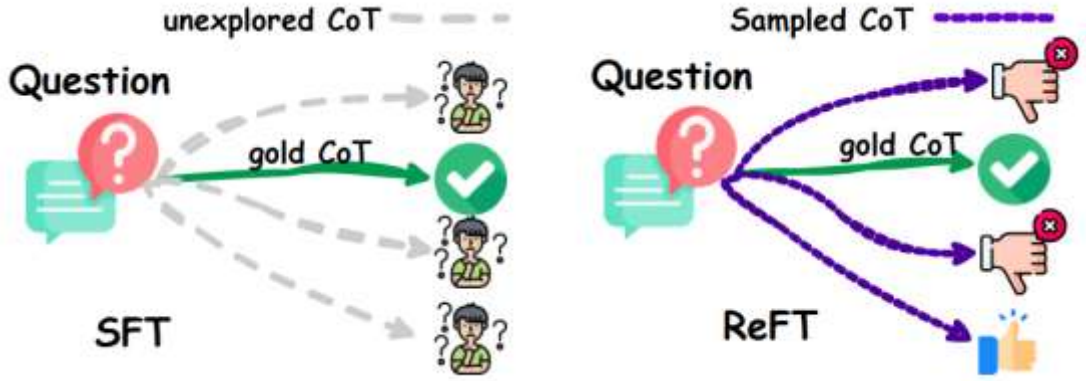
- BT model with explicit reward

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1+\exp\left(\beta\log\frac{\pi^*(y_2|x)}{\pi_{\mathrm{ref}}(y_2|x)} - \beta\log\frac{\pi^*(y_1|x)}{\pi_{\mathrm{ref}}(y_1|x)}\right)}$$

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\mathrm{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\mathrm{ref}}(y_l\mid x)}\right)\right]$$

The Secode Stage - RLVR





No human in the loop

Rule-based (Verifiable) reward:

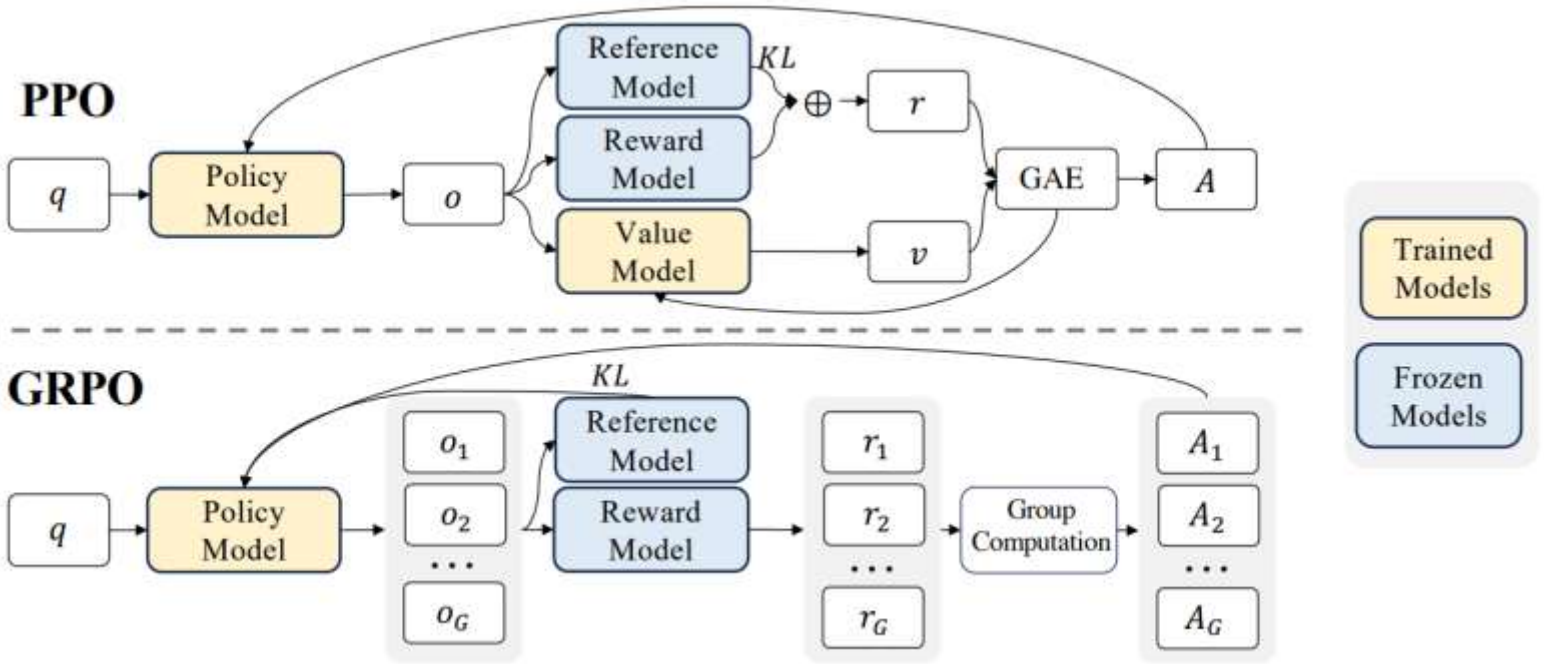
Given a rule or a gold answer,

$$r = R(s, a)$$
$$= \begin{cases} 0, & if\ a\ is\ wrong \\ 1, & if\ a\ is\ right \end{cases}$$

Format rewards:

enforces the model to put its thinking process between ‘<think>’ and ‘</think>’ tags.

Algorithm - GRPO



$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min\left[\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon\right) A_t\right]$$

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

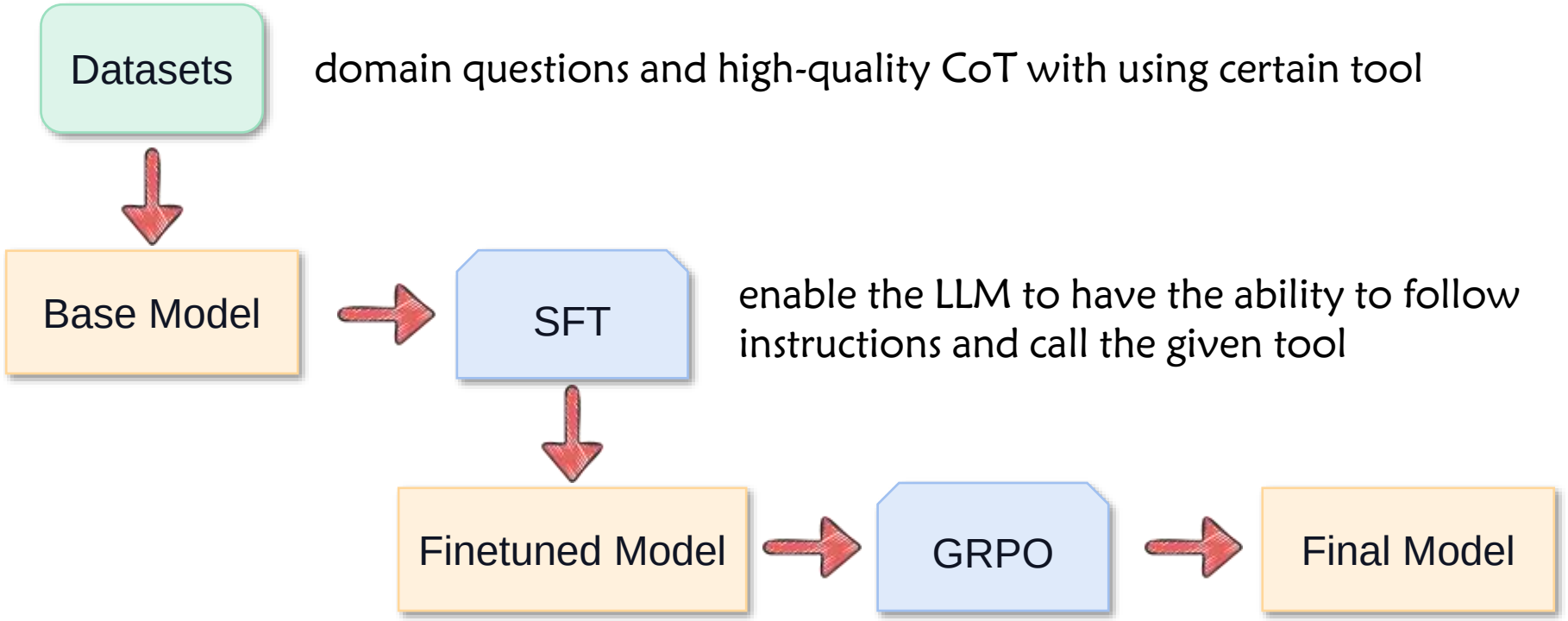
$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\}$$

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

✓ **lower memory requirements and higher computational efficiency**

Broader Applications

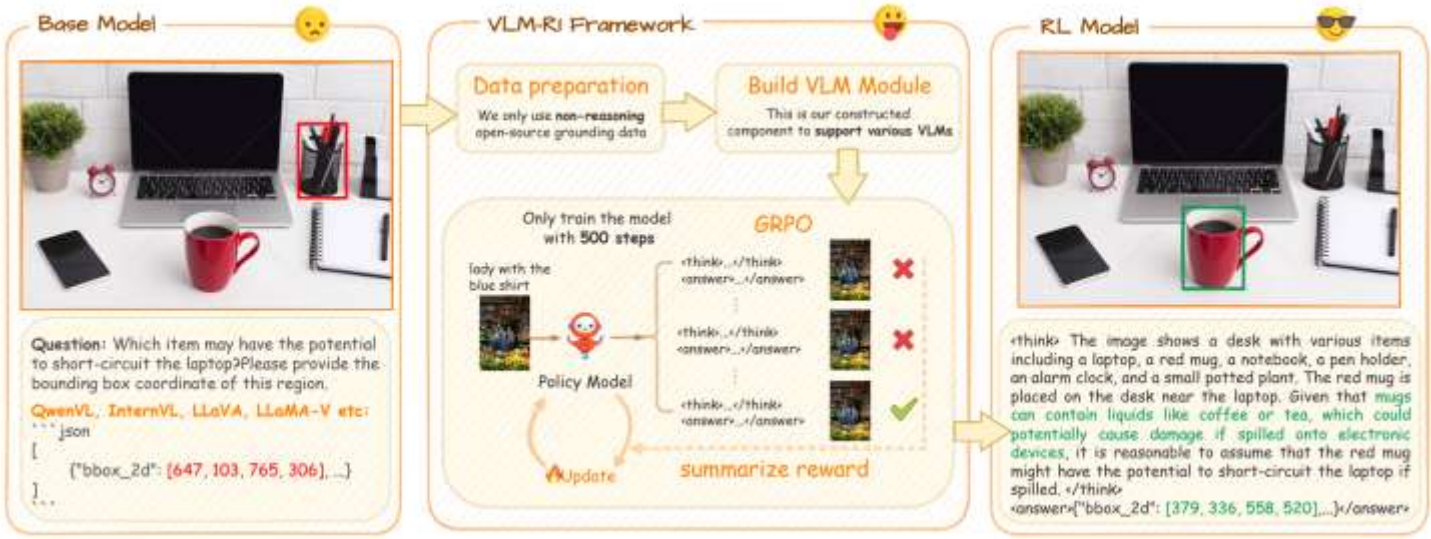
Tool-integrated reasoning



Integrate **external tools** (such as code interpreters and search engines) to enhance the reasoning ability of LLMs

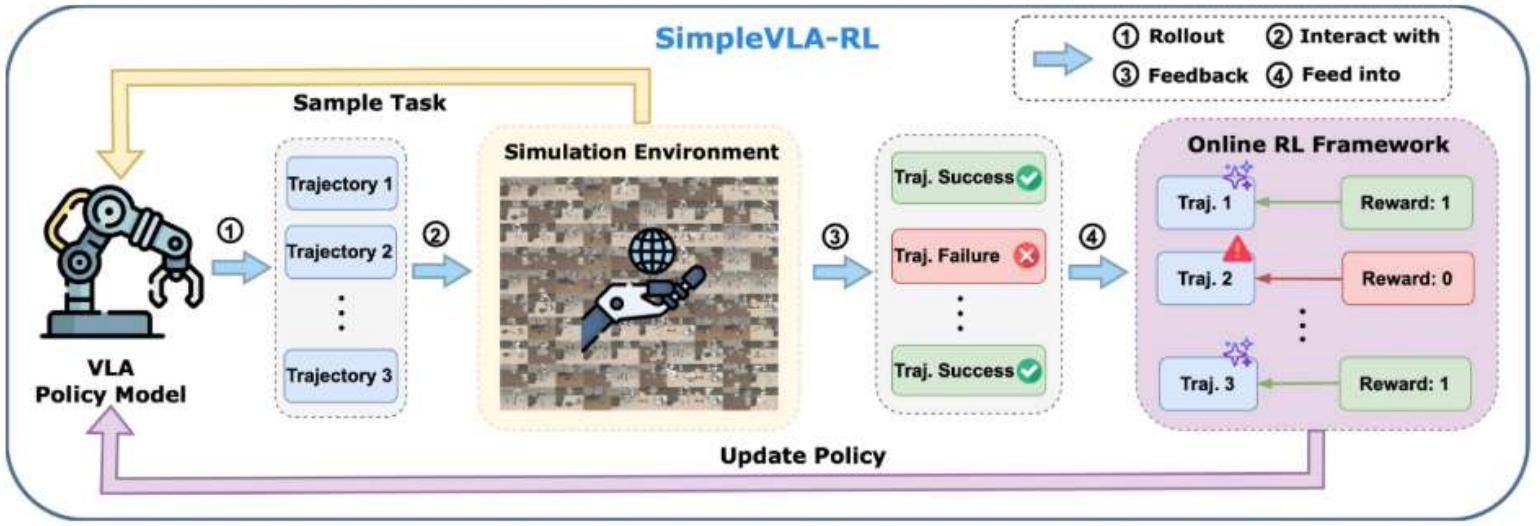
Broader Applications

VLM



Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., & others. (2025). Vlm-r1: A stable and generalizable r1-style large vision-language model. ArXiv Preprint ArXiv:2504.07615.

VLA



SimpleVLA-RL Team. (2025). SimpleVLA-RL: Online RL with Simple Reward Enables Training VLA Models with Only One Trajectory. https://github.com/PRIME-RL/SimpleVLA-RL. https://github.com/PRIME-RL/SimpleVLA-RL

The Third Stage - ?

A new generation of agents will acquire superhuman capabilities by learning predominantly from experience.

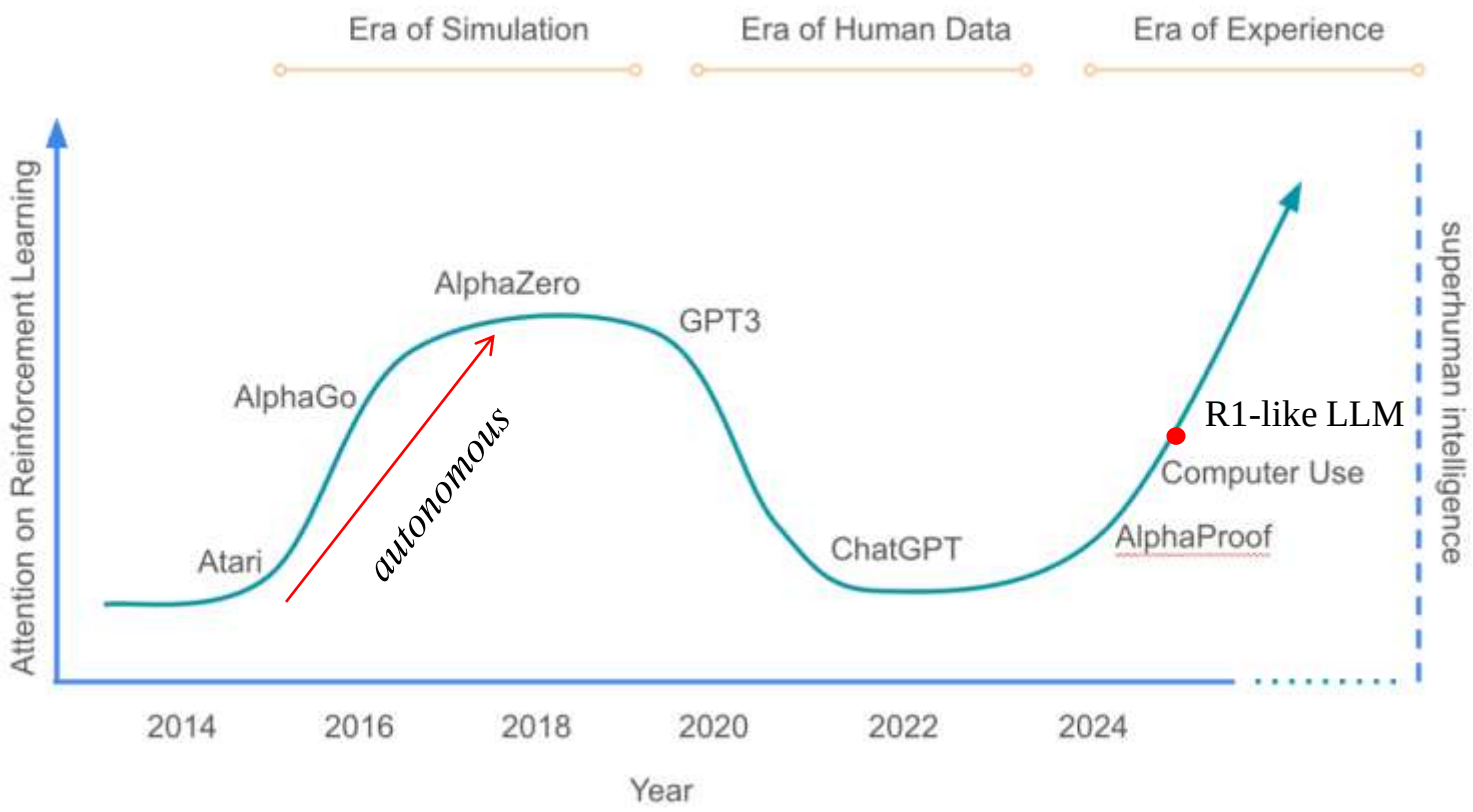
—— David Silver, Richard S. Sutton. *Welcome to the Era of Experience*

- **State Stream**: Shift from "immediate response" to "lifelong learning"

Multi-Turn interaction

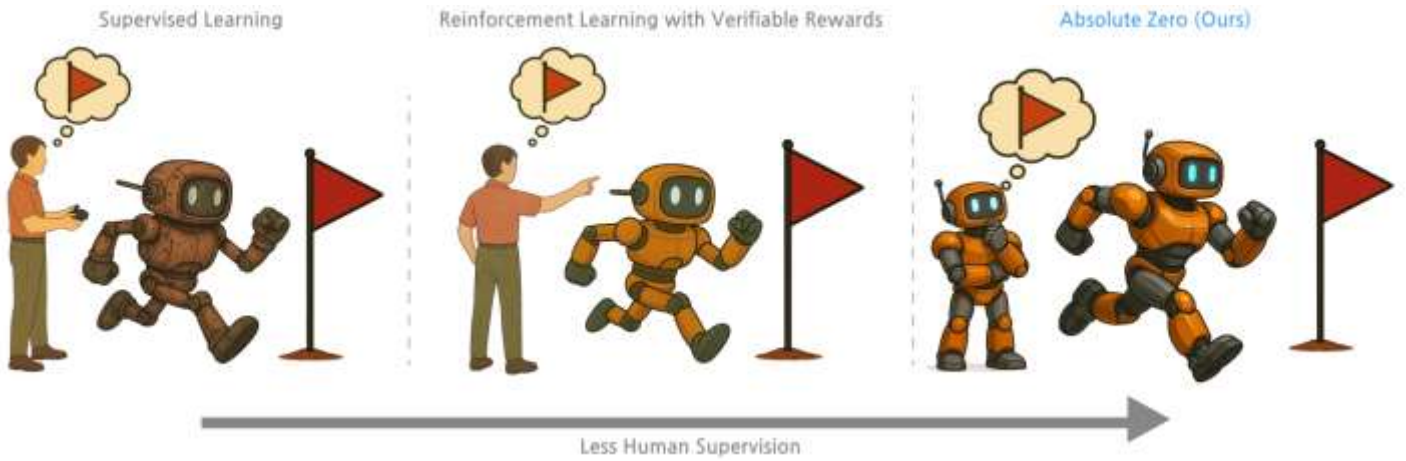
- **Actions and Observations**: Expand the scope of action to achieve autonomous exploration in the real world
- **Rewards**: A hybrid reward mechanism combining grounded signals with dynamic user feedback
- **Planning and Reasoning**: More efficient mechanisms of thought learned from interaction experience

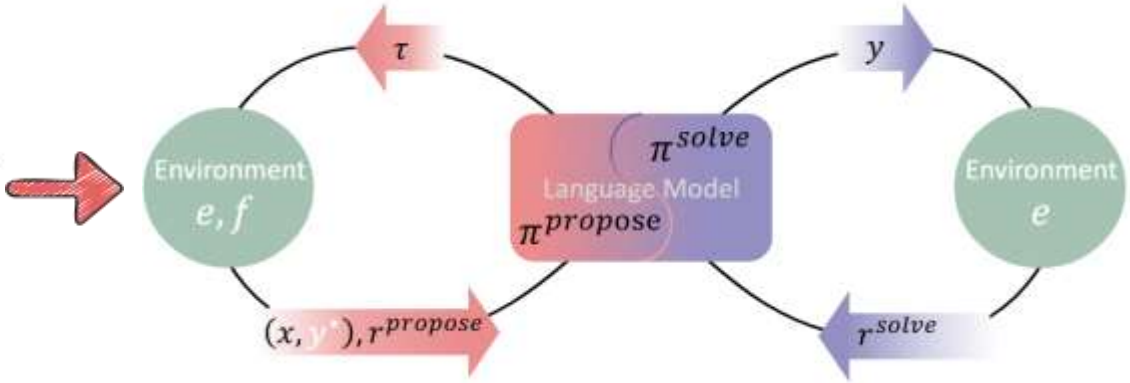
The Third Stage - ?

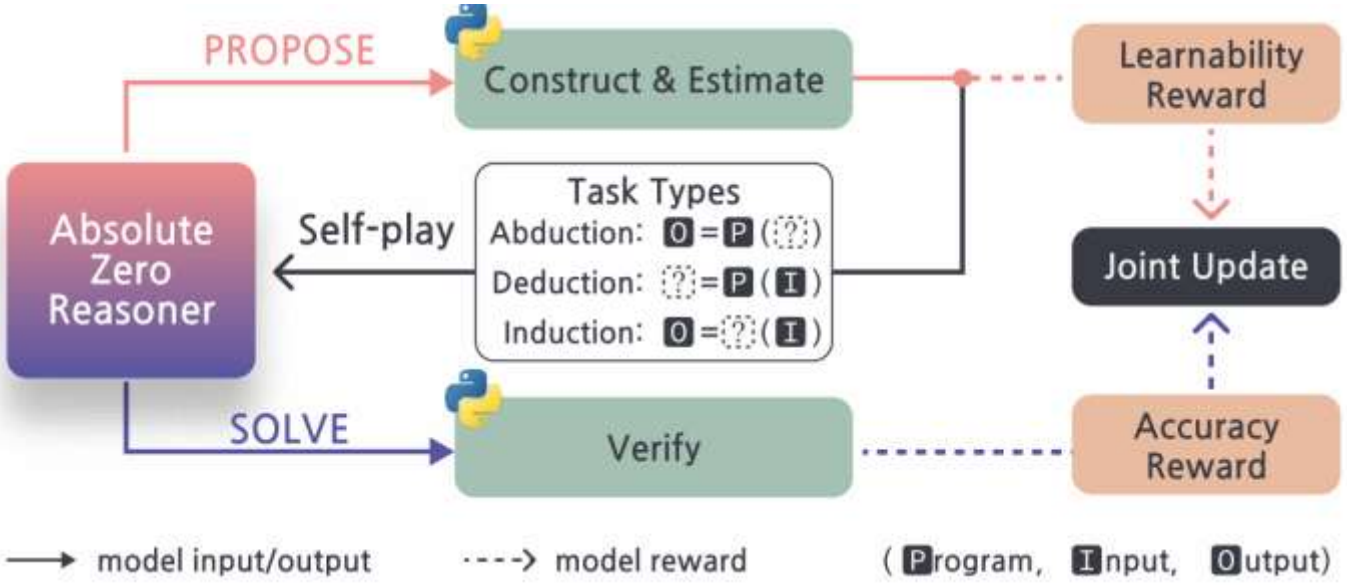


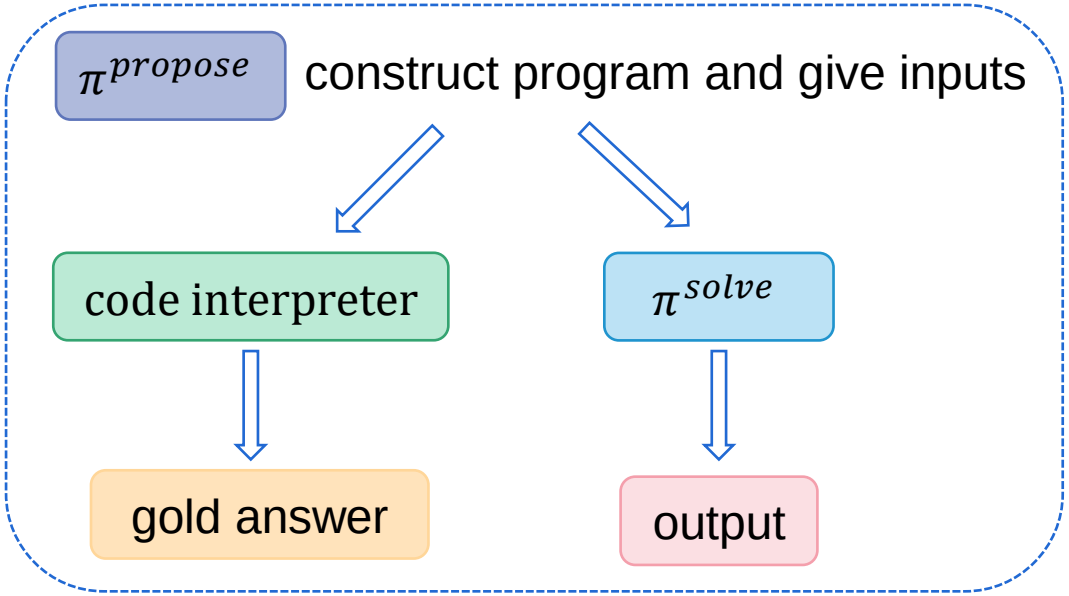
The use of massive amounts of human data as priors has finally enabled RL to generalize

Absolute Zero









Zhao, A., Wu, Y., Yue, Y., Wu, T., Xu, Q., Lin, M., Wang, S., Wu, Q., Zheng, Z., & Huang, G. (2025). Absolute zero: Reinforced self-play reasoning with zero data. *ArXiv Preprint ArXiv:2505.03335*.

Thanks